**Activity 3**

**Unsupervised-learning**

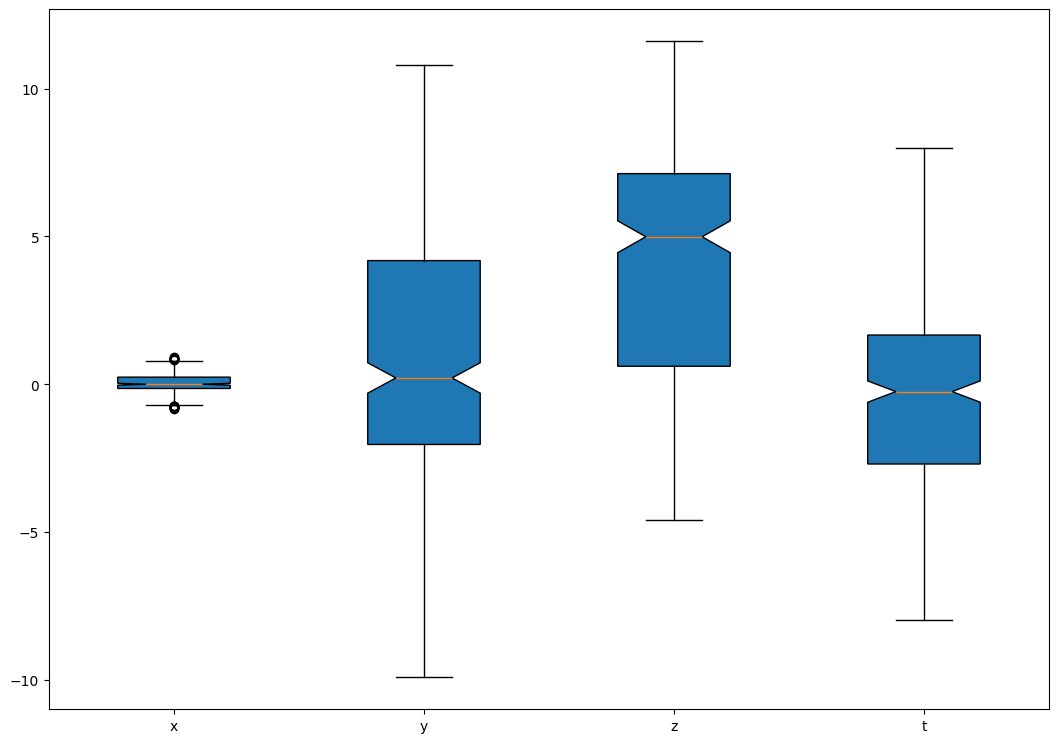
* **Git Repository**

<https://github.com/YoussefEzz/Unsupervised-learning>

* **Part 1 : Selecting and analyzing the datasets**

1. **A3-data.txt analyzed in A3-data.ipynb**

**Features**: 4 variables(**x**, **y**, **z**, **t**) of type float, 1 class of type integer(labels 1 : 5)

**Patterns**: 360 patterns

1. **2nd Dataset: Pumpkin\_Seeds\_Dataset analyzed and preprocessed in pumpkin\_seeds.ipynb**

**URL :** [**https://www.kaggle.com/datasets/muratkokludataset/pumpkin-seeds-dataset**](https://www.kaggle.com/datasets/muratkokludataset/pumpkin-seeds-dataset)

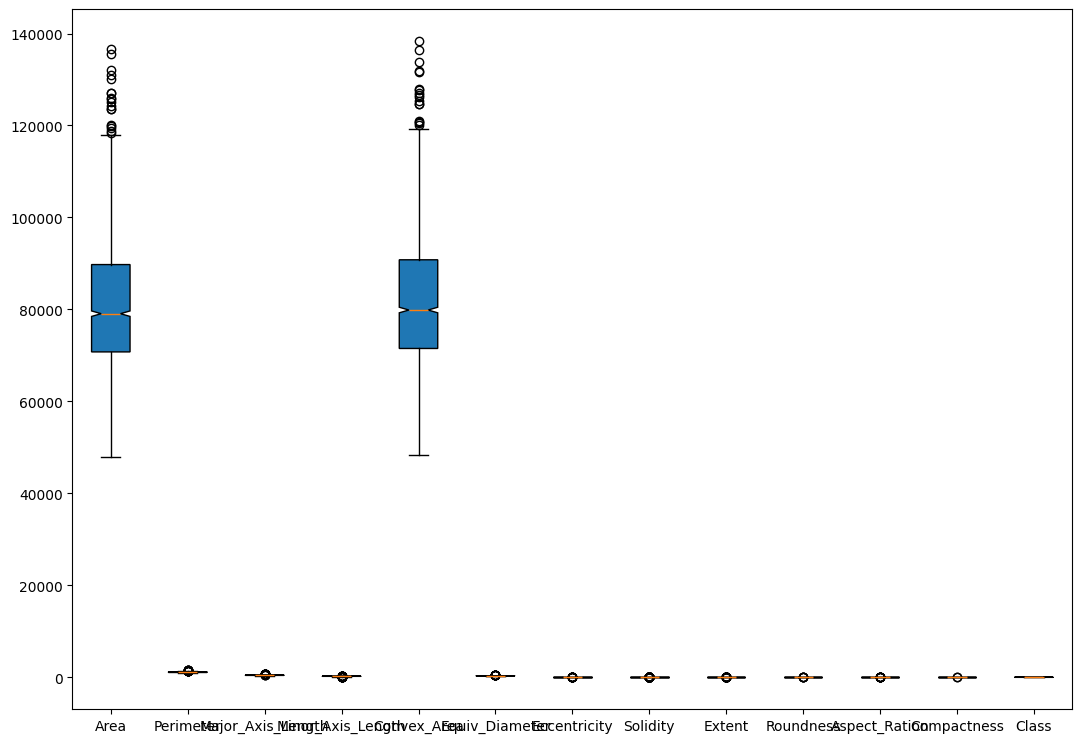
**Features**: 12 variables(Area, Perimeter, Major\_Axis\_Length, Minor\_Axis\_Length, Convex\_Area ,Equiv\_Diameter, Eccentricity, Solidity, Extent, Roundness, Aspect\_Ration, Compactness) 10 of type float and 2 of type integer, 1 class of type integer(labels 0,1) after preprocessing

**Patterns**: 2500 patterns

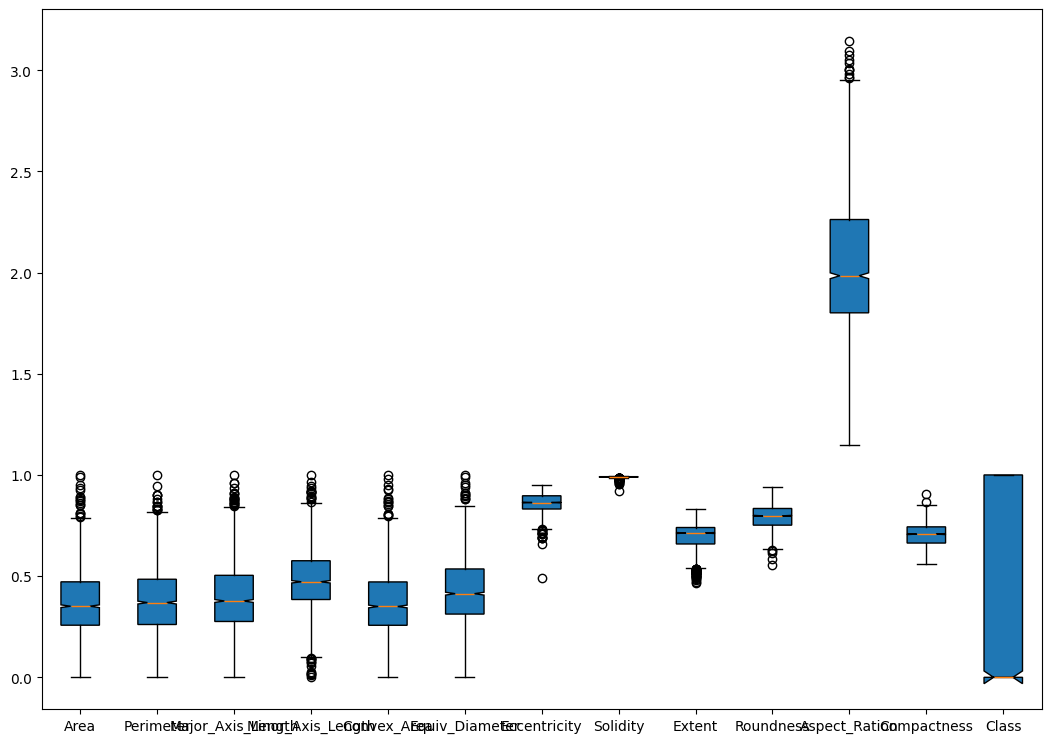
**Preprocessing**:

* + Class labels **Çerçevelik** and **Ürgüp Sivrisi** changed using label encoding to 0 and 1 respectively
  + Columns ("Area", "Perimeter", "Major\_Axis\_Length", "Minor\_Axis\_Length", "Convex\_Area", "Equiv\_Diameter") scaled to be from 0 to 1

**Before Preprocessing**

****

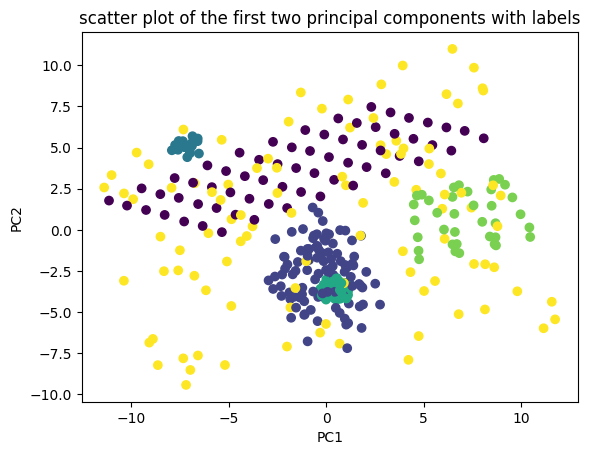
**After Preprocessing**

****

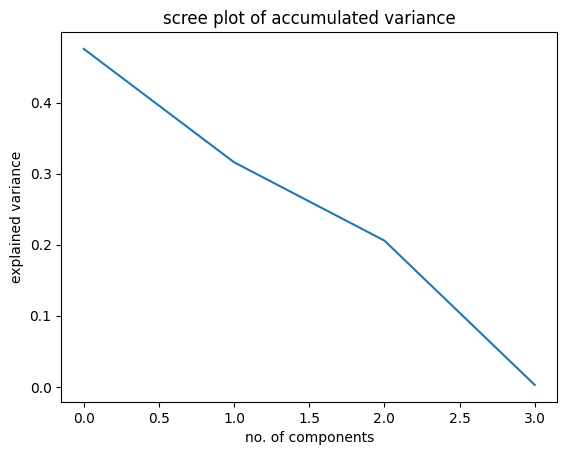
* **Part 2: Comparing unsupervised learning algorithms for A3-data.txt**

1. **PCA** analysed in **PCA.ipynb**

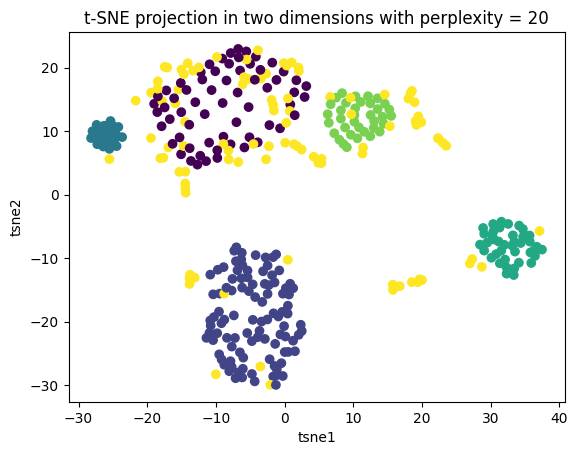
a colored scatter plot of the first two principal components

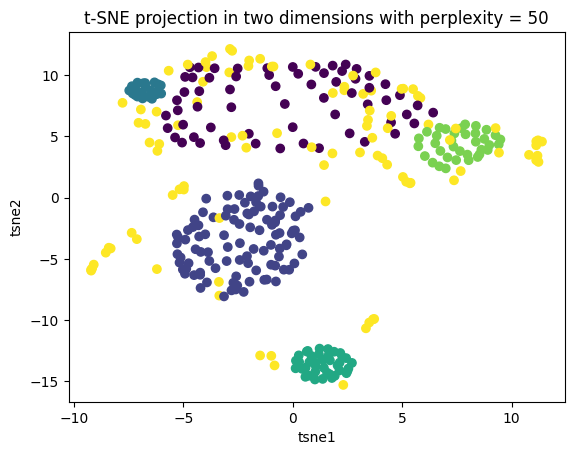


Scree plot of explained variance and no. of principal components

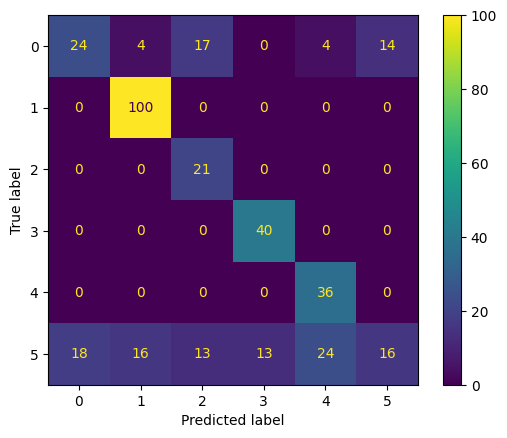


1. **t-SNE** analysed in **t-SNE.ipynb**

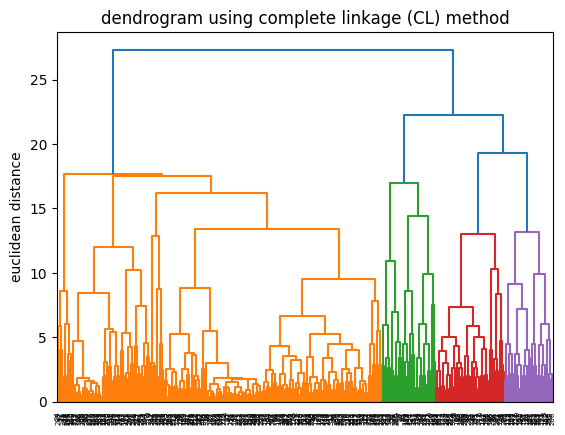
****

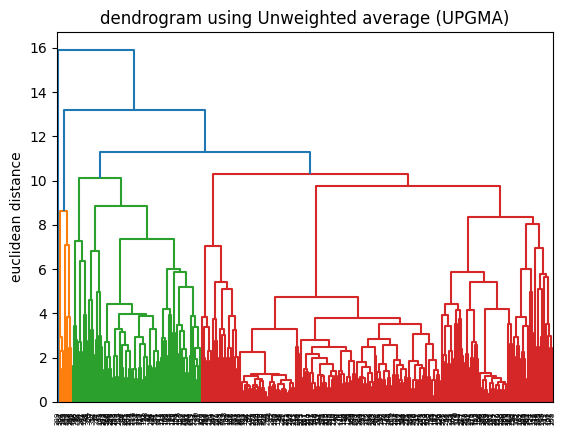
****

1. **k-means** analysed in **kmeans.ipynb**

****

1. **AHC** analysed in **AHC.ipynb**

****

****

1. **SOM** analyzed in **SOM.ipynb**

ObtainedTransformed data of shape (n, self.n\*self.m). The Euclidean distance from each item in X to each cluster center.